

An introduction in crystal structure solution and refinement

Peter Zwart

PHZwart@lbl.gov

Outline

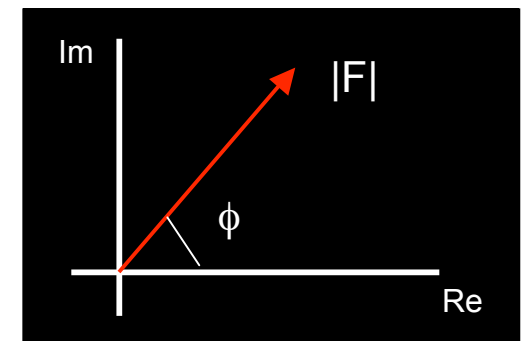
- Introduction
- Structure solution methods
 - Molecular placement
 - Molecular replacement
 - Experimental phasing
 - Direct methods
- Phase improvement
- Model building
- Refinement
- Maps

Introduction

- After collecting diffraction data and reducing it, you end up with a list of Miller indices (**H**) and intensities (**I**)
 - Intensities are the square of the structure factor amplitudes F
 - The structure factor itself is a complex quantity
 - We know its length, but do not know its ‘phase’
 - The phase is needed to compute the electron density

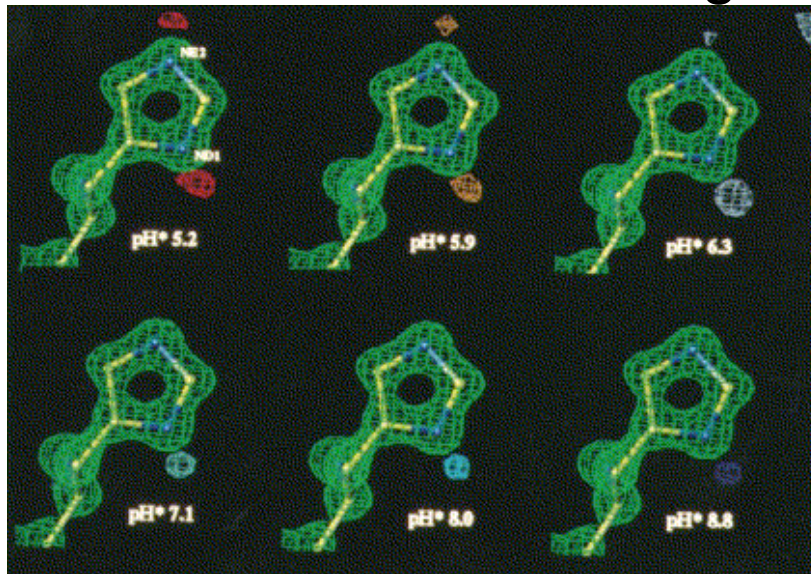
$$I_{\mathbf{h}} \propto \mathbf{F}_{\mathbf{h}} \mathbf{F}_{\mathbf{h}}^* = |\mathbf{F}_{\mathbf{h}}|^2$$

$$\rho(\mathbf{x}) = \sum_{\mathbf{h}} |\mathbf{F}_{\mathbf{h}}| \exp[-i\phi_{\mathbf{h}}] \exp[-2\pi i \mathbf{h} \mathbf{x}]$$



Introduction

- The electron density is interpreted with an atomic model
 - a collection of atoms and bonds associating them
 - When the quality and amount of data is sufficient, the level of detail can be intriguing



Berisio et al (1999)
J. Mol. Biol. 292, 845-854.

Introduction

- The measured intensities contain a wealth of structural information
- How to obtain the structure that correspond to the given data set ?
- Crystal structure determination is an iterative two stage procedure
 - Obtaining a rough guess of the phases by using the best model available. Improve and extend the atomic model by checking the electron density maps
 - Model building
 - Changing the parameters of the model so that it fits best to the data
 - Refinement
 - Iterate these steps
- How to get the initial phases though?
 - Phase problem

The solution to the phase problem

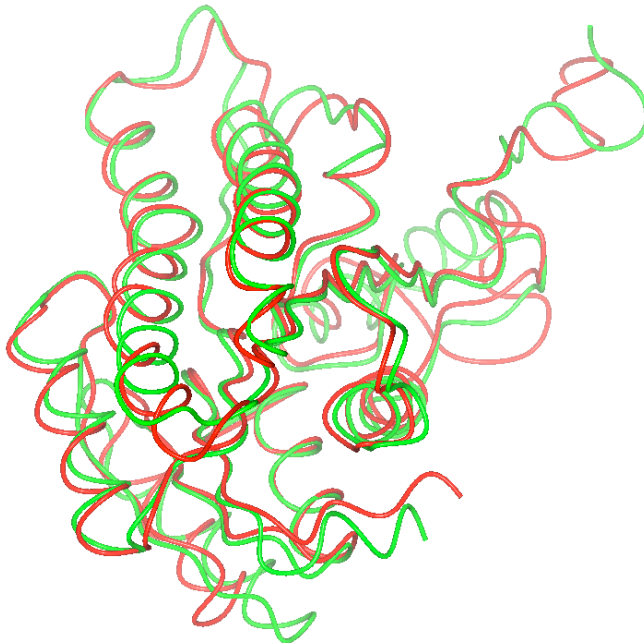
- You already have a very reasonable model
 - Protein model known in this unit cell and space group, only minor difference due to bound ligands,
 - You can start refining and looking at your maps straight away!
- You have a not so reasonable model
 - But good enough as judged from the sequence identity
 - You need to position your homologue protein in the unit cell associated with the diffraction data (molecular replacement)
- You do not have any idea how the structure looks
 - You need high resolution data or 'heavy atom' derivatives (Direct methods or experimental phasing)

“Molecular placement”

- You already have a very reasonable model
 - Protein model known in this unit cell and space group, only minor difference due to bound ligands
- The data you collected comes from a protein structure that has previously been crystallized under similar conditions
- It's unit cell and space group in the new data are very close to what it was previously
 - The model you have is probably good enough as an initial starting point.
 - No ingenuity required: you can start refinement straight away!

Molecular Replacement

- Molecular replacement utilizes structural homology between related proteins to get an initial idea of the phases



```
>>PDB:1GEE A mol:protein length:261  GLUCOSE 1-DEHYDROGE (261 aa)
  initn: 441 initl: 152 opt: 450  Z-score: 511.3 bits: 102.2 E(): 1.1e-21
  Smith-Waterman score: 450; 35.039% identity (61.811% similar) in 254 aa overlap

      10      20      30      40      50
Sequen  NDLGKTVIITGGARGLGAEMARQAVAGARVLAIVLDEEGAATA----RELGDAAAR
      :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: ::
PDB:1G  MYKDLEGKVVVITGSSSTGLGKSMAIRFATEKAKVVVNYRSKEDANSVLEIKKVGGEAI
      10      20      30      40      50      60

      60      70      80      90     100     110
Sequen  YQHLDTIEEDWQRVVAYAREEPGSVDGLVNNAGISTGMFLETSEVERFRKVVVEINLTGV
      . :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: :: ::
PDB:1G  AVKGDVTVESDVINLVQSAIKEPFGKLDVMINNAGLENPVSSHEMSLSDWNKVIDTNLTGA
      70      80      90     100     110     120

      120     130     140     150     160     170
Sequen  FIGMKTIVIPA-MKDAGGGSIVNISSAAGLMGLALTSSYGASKWGVVRLSKLAAVELGTDR
      : : . : . : . : . : . : . : . : . : . : . : . : . : . : . : . :
PDB:1G  FLGSREAIKYFVENDIKGTVINMSSVHEKIPWPLPVHYAASKGGMKIMTETLALAYAPKG
      130     140     150     160     170     180

      180     190     200     210     220     230
Sequen  IRVNSVHPGMIYTPMTAS--TGIRQGGEGNYPTFMGRVGPGEIAGAVVKLLSDTSSYVT
      : : : : . : . : . : . : . : . : . : . : . : . : . : . : . :
PDB:1G  IRVNNIGPGAINTPINAERFADPEQRADVESMIFMGYIGEPERIAAVAAVLASSEASYVT
      190     200     210     220     230     240

      240     250
Sequen  GAELAVDGGWTTGPTVKYVMGQ
      : : . : . : . : .
PDB:1G  GITLFADGGMTLYPSFQAGRG
      250     260
```

Molecular Replacement

- The solution strategy is to take the model you think looks most like the protein structure of interest, and place it in the unit cell
 - Use sequence alignment tools to find a template for your molecule
- In most cases, you need to determine 6 parameters
 - 3 parameters describing the orientation
 - 3 parameter describing the location
 - A six dimensional search is very time consuming
- As it turns out, you can split the search into two different sub problems:
 - Rotation function to find the orientation
 - Translation function (with a fixed orientation) to find the location

The Patterson Function

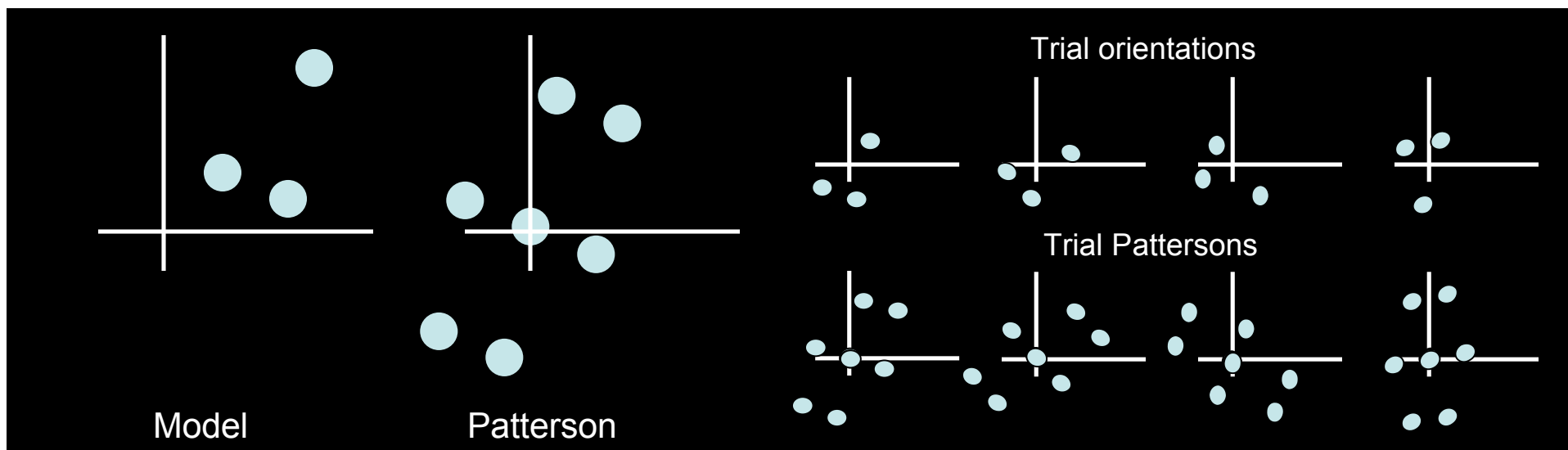
- The Patterson function can be computed from the experimental data
 - No phase information is needed
- The Patterson function is a 3 dimensional 'map' with maxima corresponding to inter atomic vectors
 - Huh?
 - If you have an atom at \mathbf{x}_1 and \mathbf{x}_2 , The Patterson function will have peaks at
 - $0,0,0$ ($\mathbf{x}_1-\mathbf{x}_1$; $\mathbf{x}_2-\mathbf{x}_2$)
 - $\mathbf{x}_1-\mathbf{x}_2$
 - $\mathbf{x}_2-\mathbf{x}_1$
 - $\mathbf{x}_1-(R\mathbf{x}_1 + \mathbf{T})$ (symmetry related peaks)
 - $\mathbf{x}_1-(R\mathbf{x}_2 + \mathbf{T})$ (symmetry related peaks)
 -

The Patterson Function

- The origin peak of the Patterson is due to interatomic vectors to itself
 - And because there are lots of those, this peak is really big
- The vector length of the location of Patterson peak is equal to the inter atomic distance
 - The area of the Patterson close to the origin is mostly populated by inter atomic vectors from atoms within a molecule
 - Further away from the origin you get inter atomic vectors from atoms in different (possibly symmetry related) molecules

The Rotation Function

- The rotation function determines the orientation of the search model in the unit cell of the crystal structure under investigation
- 3 parameters need to be determined
- The basis of the rotation function lies in the Patterson function
 - Modern implementations of the rotation function involve rather complex mathematics, mostly based on spherical harmonics (brrrr)
 - A 'real space' version is however easy to understand



The Translation Function

- The translation function describes the fit of a molecule to the data as a function of its position in the unit cell
- It can be computed relatively fast (FFT's are involved)
- Various scoring functions are possible
 - CC on I (AMORE, MOLREP)
 - CC on F (AMORE, MOLREP)
 - Likelihood (PHASER)

The Translation Function

- For each rotation function solution, a translation function has to be computed
 - If the solution to the rotation function is ambiguous, you end up calculating a lot of translation function
 - This can get complicated and costly when you are looking for multiple copies in the ASU
 - Good book keeping is essential
 - PHASER does an excellent job here

Experimental phasing

- Sometimes molecular replacement will not work and other approaches are needed
- Experimental phasing is the only alternative
 - in 99% of the cases at least
- Experimental phasing relies on the introduction of 'heavy atoms' in crystal
- Two routes
 - Isomorphous replacement (SIR , MIR)
 - Anomalous scattering (SAD , MAD)

Isomorphous replacement

- For isomorphous replacement, two (or more) data sets are needed
 - The protein
 - The protein with a bound heavy atom (Hg, Au, Pt, Br, I, ...)
- Differences in intensities (isomorphous differences) of the two data sets is fully ascribed to the presence of the heavy atoms
 - Since there are not many heavy atoms, and the unit cell is quite large, a isomorphous difference Patterson function can be used to find the sites
- The location of the heavy atom and the two amplitudes (F_{nat} and F_{der}) can be enough to get a reasonable estimate of the phase of F_{nat}
 - More independent derivatives give better estimates in theory
 - This need not be in practice though

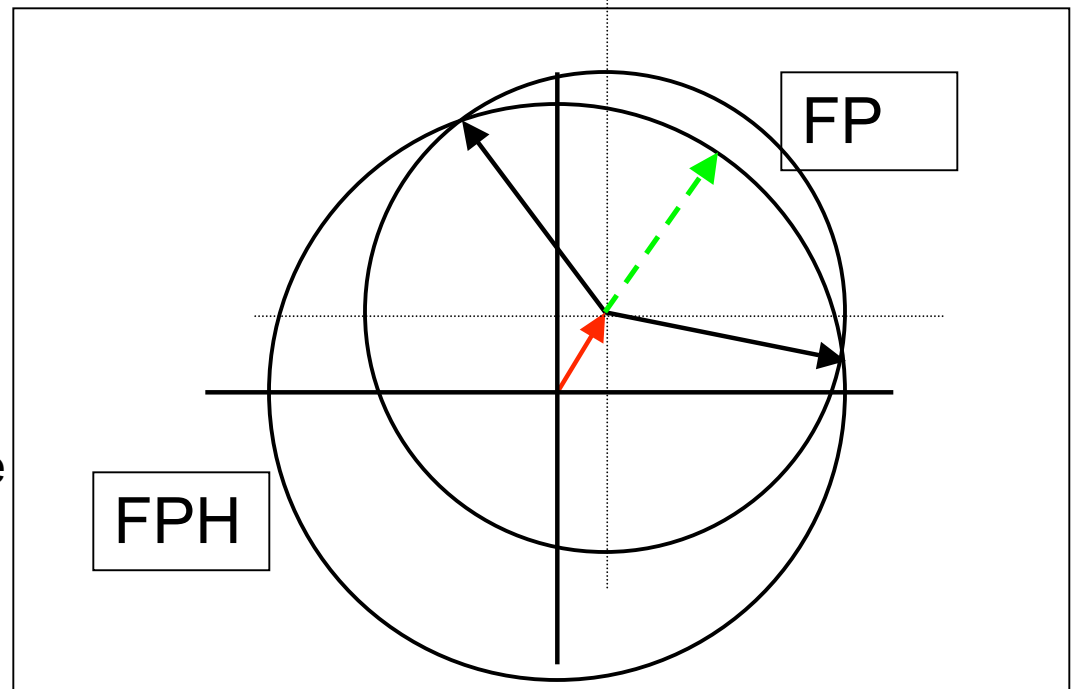
Isomorphous replacement

- For isomorphous replacement, two (or more) data sets are needed
 - The protein (FP)
 - The protein with a bound heavy atom (Hg, Au, Pt, Br; FPH)

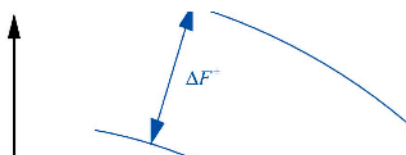
From two amplitudes and a heavy atom position, two phase choices can be obtained (phase ambiguity)

The average of those is a good start

A third data set would nail the phase down unambiguously



Anomalous scattering

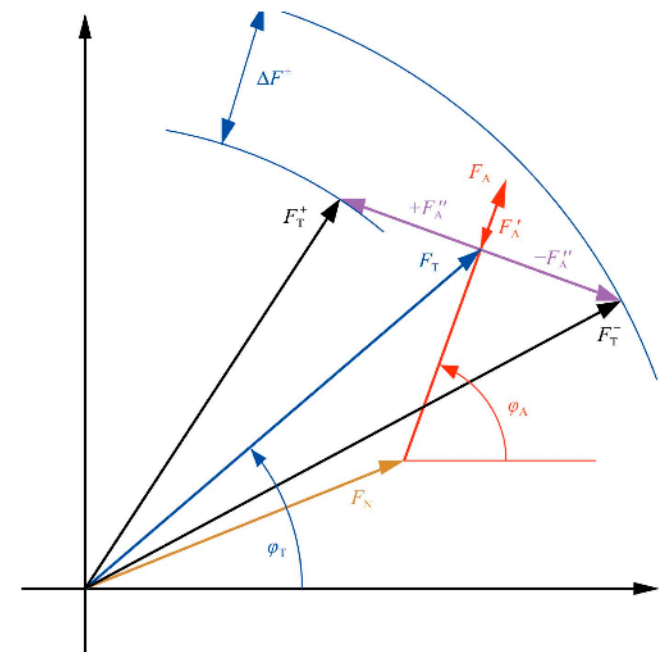
- If the incident radiation on a crystal is close to an absorption edge of an atom that is in the structure, 'funny' things start happening
 - The 'form factor' is a complex quantity
 - $f_{\text{tot}} = f^0 + f' + if''$
 - f' and f'' depend on wavelength
- 

$$F_{\mathbf{h}} = \sum_j (f_j^0 + f_j' + if_j'') \exp[-2\pi i \mathbf{h} \mathbf{x}_j]$$

$$F_{-\mathbf{h}} = \sum_j (f_j^0 + f_j' + if_j'') \exp[2\pi i \mathbf{h} \mathbf{x}_j]$$

$$F_{-\mathbf{h}}^* = \sum_j (f_j^0 + f_j' - if_j'') \exp[-2\pi i \mathbf{h} \cdot \mathbf{x}_j]$$

- $|F_h|$ not necessarily equal to $|F_{-h}|$



Wang et al, Acta Cryst D63, 751-758 (2007)

Anomalous scattering

- Under 'normal' circumstances, Friedel's law holds:

$$I_{\mathbf{h}} = I_{\bar{\mathbf{h}}}$$

- When the 'heavy' atoms are present and the wavelength is close to the absorption edge, Friedel's law doesn't hold

$$I_{\mathbf{h}} \neq I_{\bar{\mathbf{h}}}$$

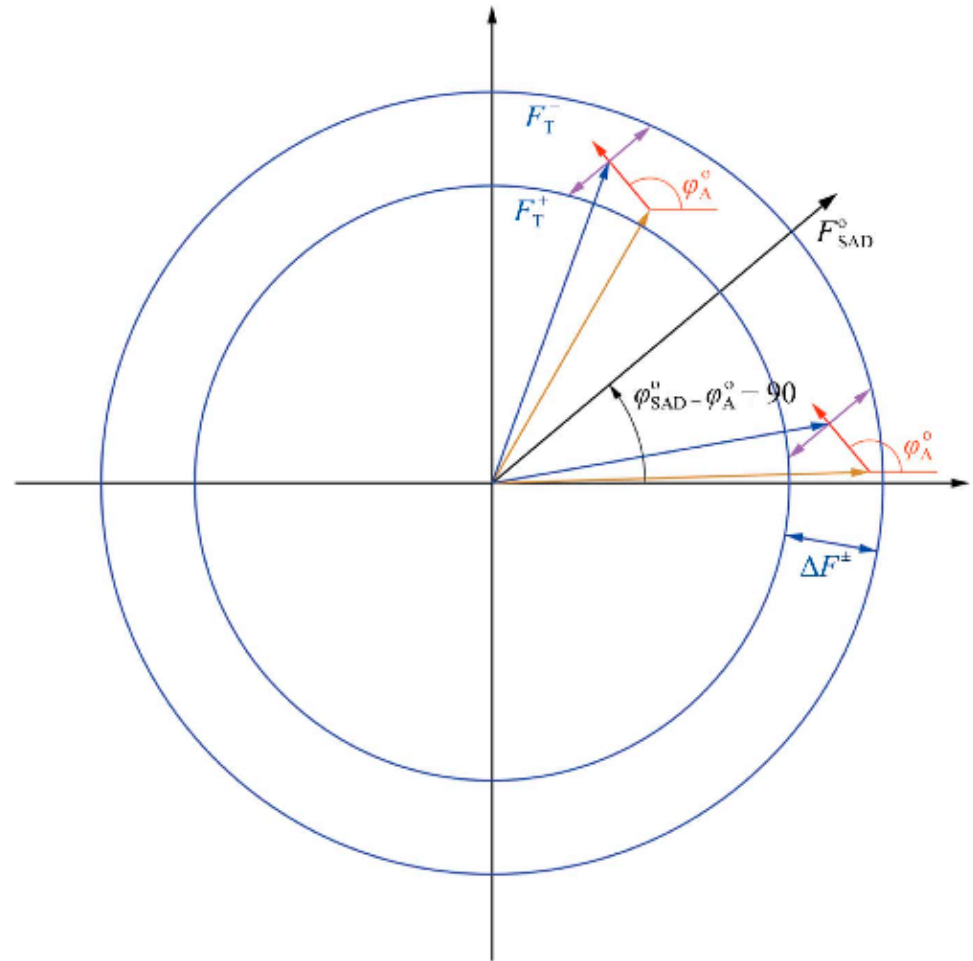
- The anomalous differences are approximately proportional to the amplitude of the heavy atom structure that is causing it:

$$|F_{\text{heavy}}| \propto \left| |F^+| - |F^-| \right|$$

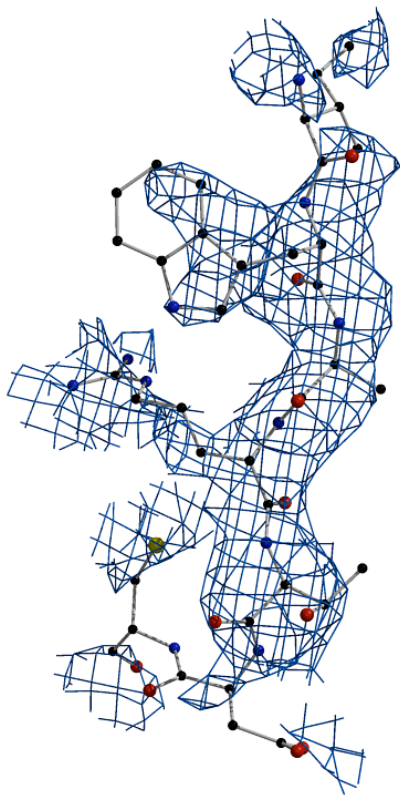
- Patterson methods can be used to find the sites

SAD Phasing

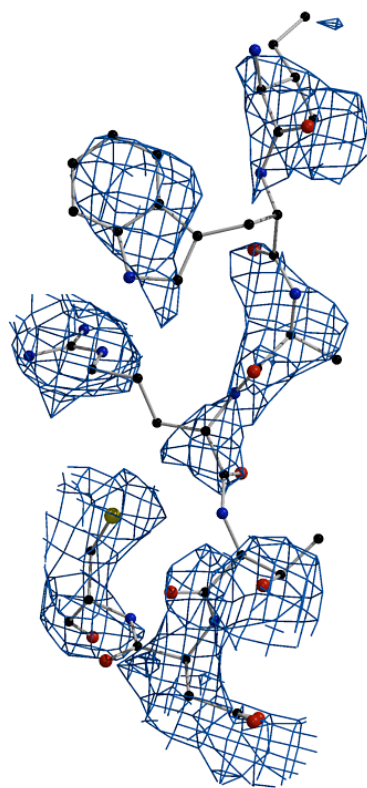
- **Single-wavelength Anomalous Diffraction**
 - Again two phases are possible, one of them is more likely than the other
 - With a one more wavelength (MAD), you would lose the ambiguity
 - In theory



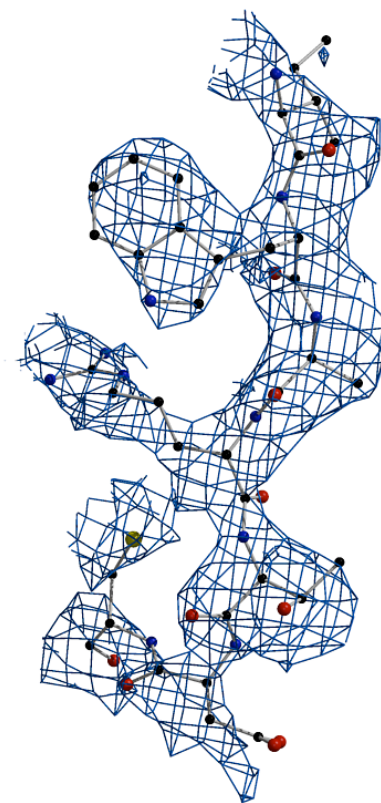
SIRAS



ISO



ANO



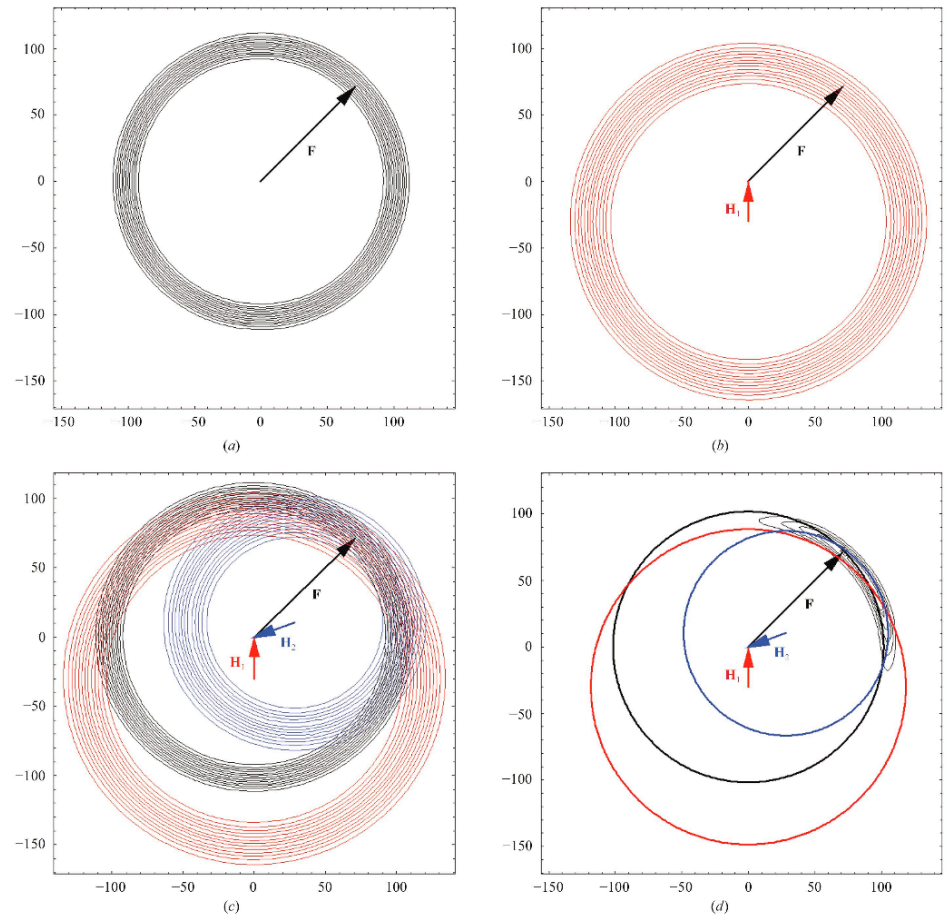
ISO+ANO

In an ideal world

- With no experimental errors, a SAD experiment will give you an average cosine of the phase error over the whole acentric data set that is close to 0.60
 - Even if the you only has 1 single Sulfur in 50000 residues
 - Due to pure geometry
 - The 53 degrees can be readily improved via solvent flattening
- Under similar circumstances, MAD will give you phases that have no errors
- Similar arguments for SIR(AS) / MIR(AS)

In reality however

- We do have errors
 - Counting statistics
 - Errors introduced during integration and scaling
 - Radiation damage
 - Gradual introduction of non-isomorphism to 'itself'
 - Non isomorphism between native and derivative
 - 'Correlated non-isomorphism' between derivatives
- A proper statistical treatment is needed to handle errors appropriately
- Increasing number of datasets/derivatives does not necessarily result in better phases



Direct methods

- Direct methods is a class of solution techniques that generates good starting phases using only experimental intensities as a source of phase information
- The basis of direct methods are (in most cases)
 - Approximately equal atoms
 - Non-negativity of the electron density
 - Atomicity of density
 - a few well-defined, non overlapping peaks

Direct methods

- When previous conditions are met, we have
$$\rho(\mathbf{x}) \approx k\rho^2(\mathbf{x})$$
- Basic structure solution scheme:
 - 0. Take random starting phases, compute map with Fobs
 - 1. Square the observed map, back transform to get new phases
 - 2. Combine phases with Fobs, compute new map
 - 3. Go to 1; Cycle until done
 - Pick peaks and find model
- Multiple random starts are needed
- Step 1 can be done more efficiently via a an expression called the *tangent formula*

Direct methods

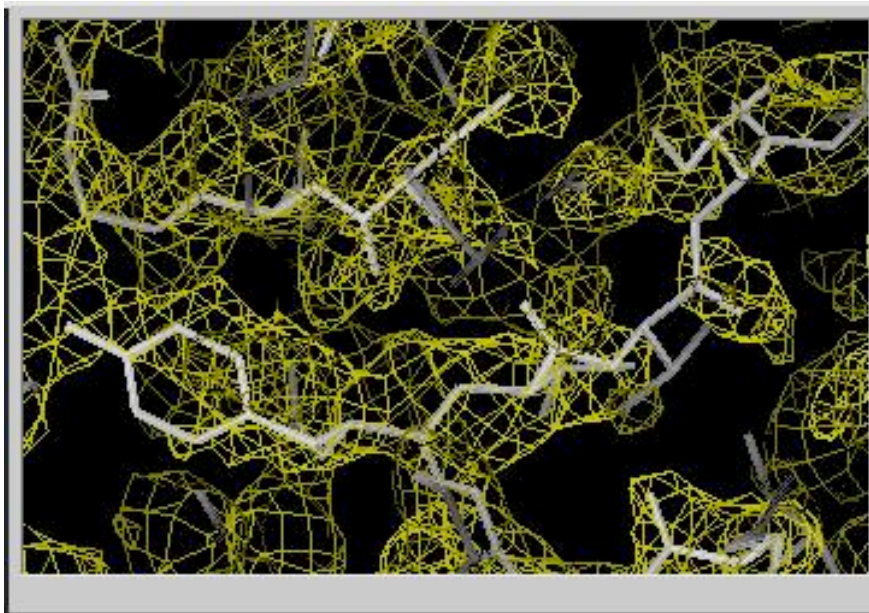
- Direct methods can be combined with Patterson techniques to get better than random phases
 - Higher success rate for each trial
- You can pick peaks in intermediate maps as well and use an atomic model to compute phases
 - Faster convergence of iterative procedure
- Not only can you solve 'regular' structures this way, but substructures as well!
 - Direct methods are now the main vehicle for solving substructures from anomalous/isomorphous data
- SnB, SHELXD and phenix.hyss use these methods

Phase improvement

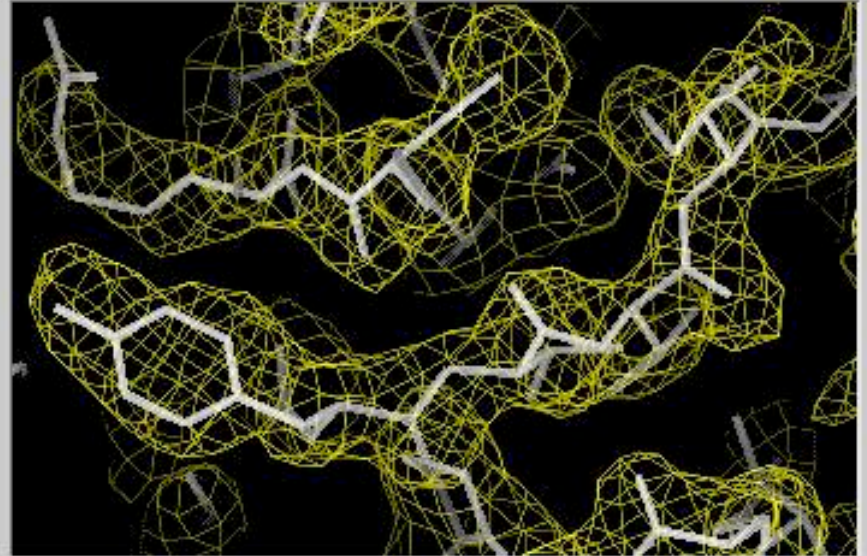
- Often, starting phases (from EP or MR) can be improved by changing the phases in such a way that certain prior knowledge about how protein electron density is satisfied.
 - Flatness of bulk solvent
 - Histogram of protein region
 - NCS relations between density
 - Very powerful
 - Relations between different crystal form
 - Very powerful
- This procedure is called density modification
 - One of the most powerful tools for improving phases when no atomic model is present

Phase improvement

- Density modification software:
 - DM, SOLOMON, RESOLVE, PIRATE



MAD phases; CC=0.37



Resolve phases; CC=0.79

Model building

- Model building can be done by hand
 - O, COOT, XtalView, TurboFRODO, MIFIT
- Model building can be done automatically
 - ARP/wARP, RESOLVE
 - It is an iterative process that mixes interpretation of density with refinement of model / phase improvement by density modification
- Automated model building can give you a complete model at when the resolution of your model is reasonable (say 2.5Å or better)
 - It also depends on the solvent content and quality of initial phases

Refinement

- Refinement is the part of the structure solution procedure where you 'finish up' your model
- The model is parameterized by atoms which have
 - Positional parameters (3)
 - Atomic displacement parameters (1, or 6)
- Besides Fobs you have a preconceived notion of bond lengths and angles: restraints
 - The restraints act as an additional set of observations

Refinement

- Refinement optimizes the function
$$Q(\text{model}) = Q(\text{data} \mid \text{model}) + Q(\text{model} \mid \text{restraints})$$
- Model has parameters
 - (x,y,x)
 - Biso (or Baniso)
 - Scale factor
- Use standard numerical techniques to change parameters of model as to improve $Q(\text{model})$

Q(model | data)

- Xray target function (or neutrons)

- Least squares on F

$$Q_{\text{lsqf}} = \sum_h w_h \left(|F_{\text{obs}}| - k |F_{\text{model}}| \right)^2$$

- Least squares on I

$$Q_{\text{lsqI}} = \sum_h w_h \left(I_{\text{obs},h} - k |F_{\text{model},h}|^2 \right)^2$$

- Likelihood on F

$$Q_{\text{mlf}} = \sum_h \log \left[P(F_{\text{model}} | F_{\text{obs}}, \sigma_A) \right]$$

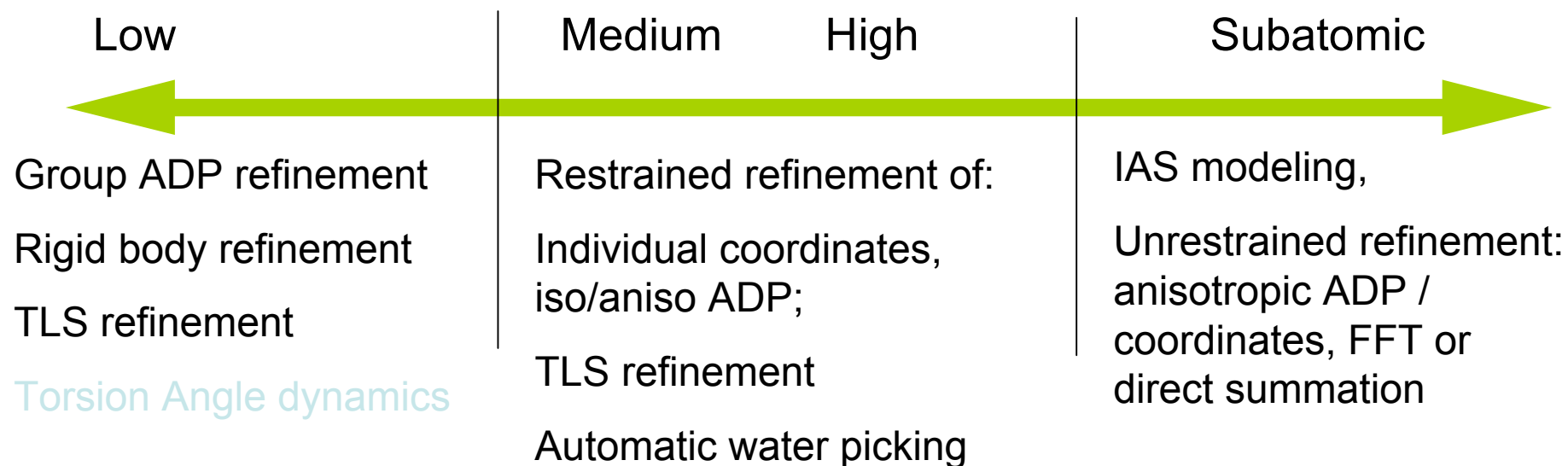
Likelihood based refinement

- Likelihood based refinement has proven to have a larger radius of convergence than least square target function
- Likelihood based refinement takes into account the current quality of the model during refinement
 - It automatically weights down data that is not supposed to fit well due to model error (high reso mainly)
 - When the model gets better, the high resolution data becomes more important
 - This variable weighting is the reason why ML refinement works well. If likelihood based weights are introduced in LS refinement, very similar results are obtained

Likelihood based refinement

- The presence of anomalous data can further enhance refinement
 - Phase probability distributions obtained from experimental phasing can be used as observations and increase the stability of the refinement
 - MLHL target
 - REFMAC, CNS, phenix.refine

Refinement strategies



Refinement strategies

- Optimization of placement of large, fixed bodies
 - Rigid body refinement. 6 parameters per domain
- Optimisation of coordinates
 - 3 parameters (or less) per atom
- Optimisation of ADP's
 - Isotropic: 1 parameter per atom (a sphere)
 - Anisotropic: 6 or less parameters (an ellipsoid)
- Occupancies
 - 1 parameter per atom/group
- f'/f''
 - 2 parameters per atom / group

Domain movement

- Sometimes large domains 'move' in a crystal
- This can be describe by a TLS model
 - 19 parameters per domain
 - Describes anisotropic movement of a domain
 - Common when ASU contains more than a single molecule
 - Has potential to reduce R values massively

Domain movement

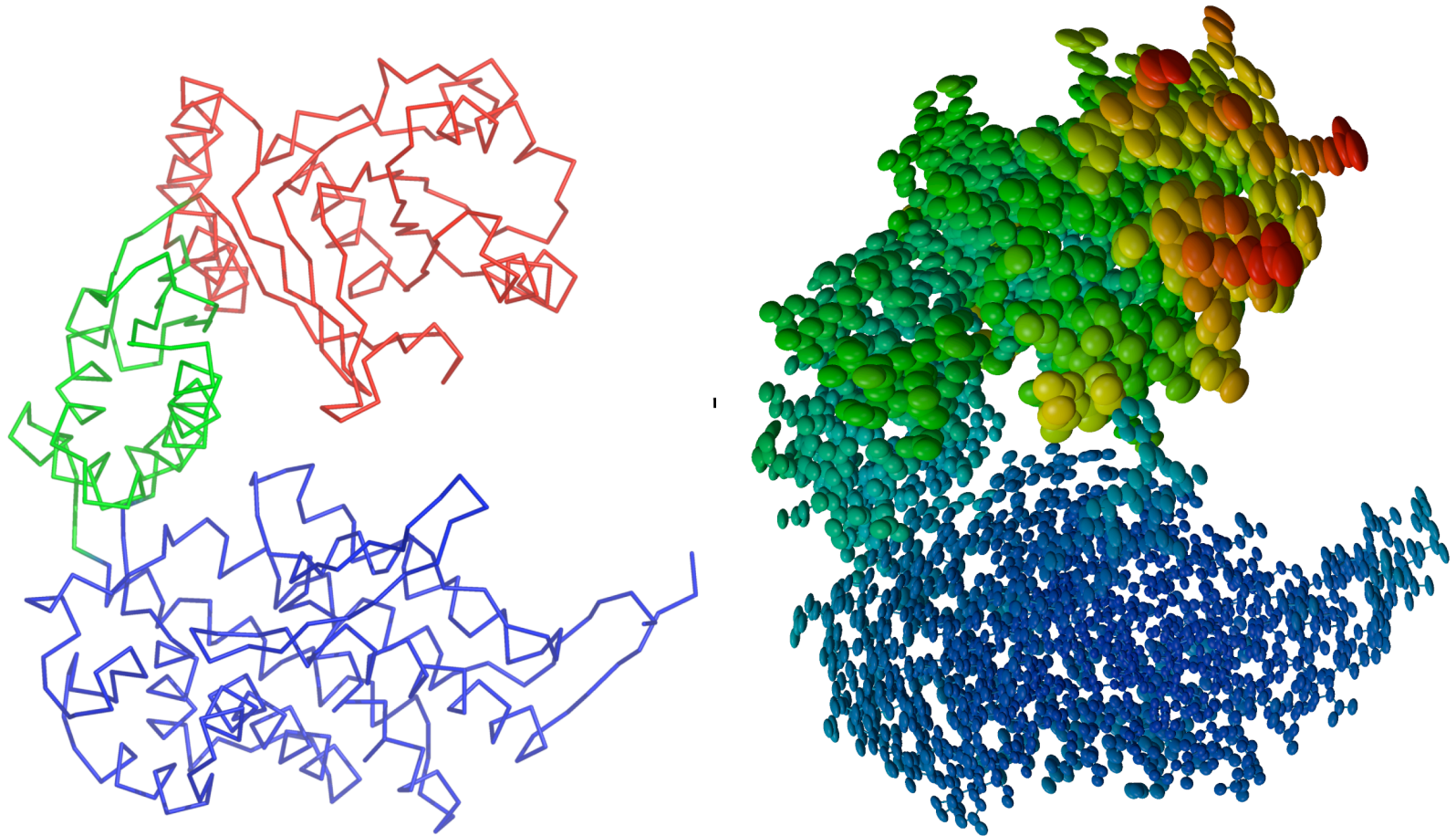


Image from Paul Adams

Refinement results from phenix.refine

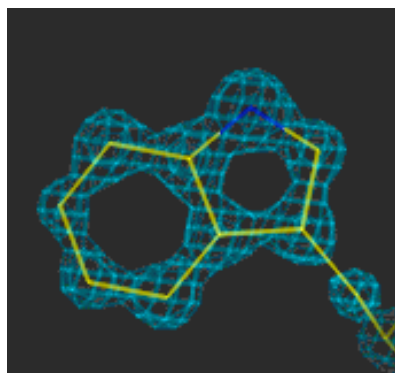
Validation of results

- Xray data:
 - R-value
 - Computed on data against which the structure is refined
 - Free R-value
 - Compute on data against which the data has not been refined
 - ‘unbiased’
 - Availability of raw data / images
 - To make sure no-one can accuse you of fabricating the structure
- Model
 - Ramachandran plot
 - Sort of ‘unbiased’
 - Clash scores and other geometry based criteria
 - Google on **MOLPROBITY** to find the site
 - More up to date validation criteria than procheck

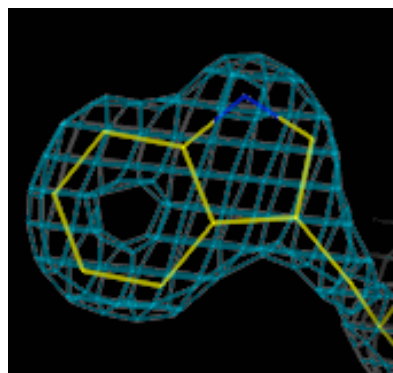
$$\frac{\sum_h |F_{\text{obs}} - F_{\text{calc}}|}{\sum_h F_{\text{obs}}}$$

Maps

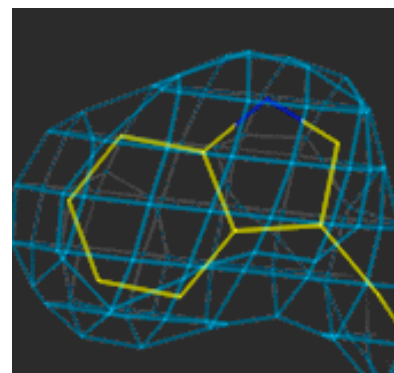
- Electron density maps describe how many electrons are sitting where in the unit cell
 - Low resolution maps do not reveal much
 - High resolution maps give loads of information



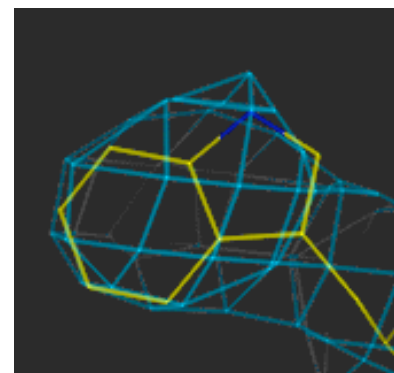
1Å



2.5 Å



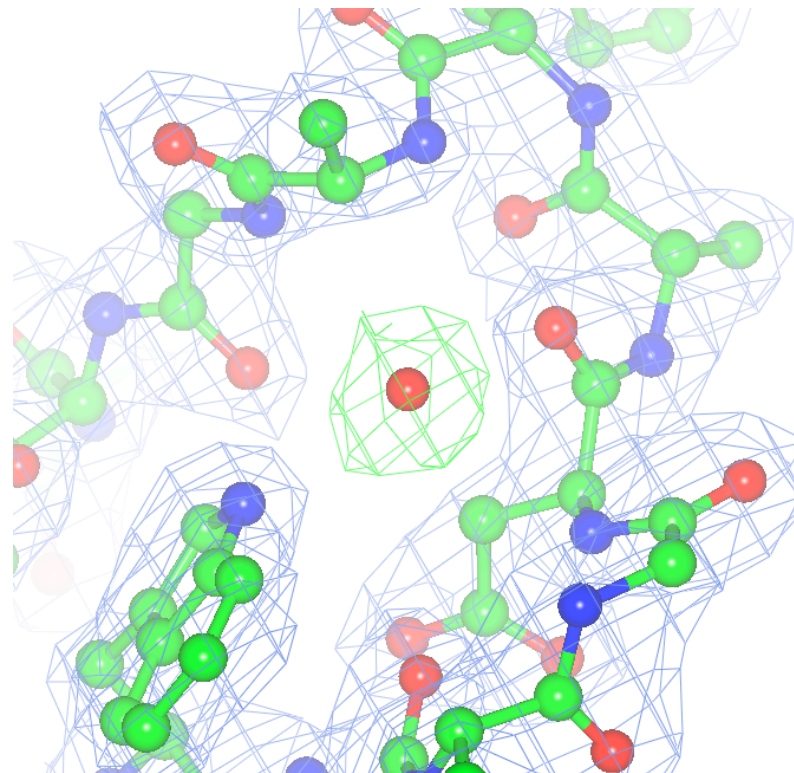
3Å



4Å

Maps

- Coefficients
 - Electron density
 - $2F_o - F_c$, PHlc
 - $(F_o, PHlc) - (F_o - F_c, PHlc)$
 - $2mF_o - DF_c$, PHlc
 - $(mF_o, PHlc) - (mF_o - DF_c, PHlc)$
 - Difference map
 - $F_o - F_c, PHlc / mF_o - DF_c, PHlc$
 - Indicates the where the current model lacks electrons (positive peaks) or has too many electrons (negative peaks)
 - m : expected cosine of the phase error
 - D : The fraction of F_{calc} that is correct
 - M and D are correlated and estimated by a simple numerical procedure
 - sigmaA estimation



Maps

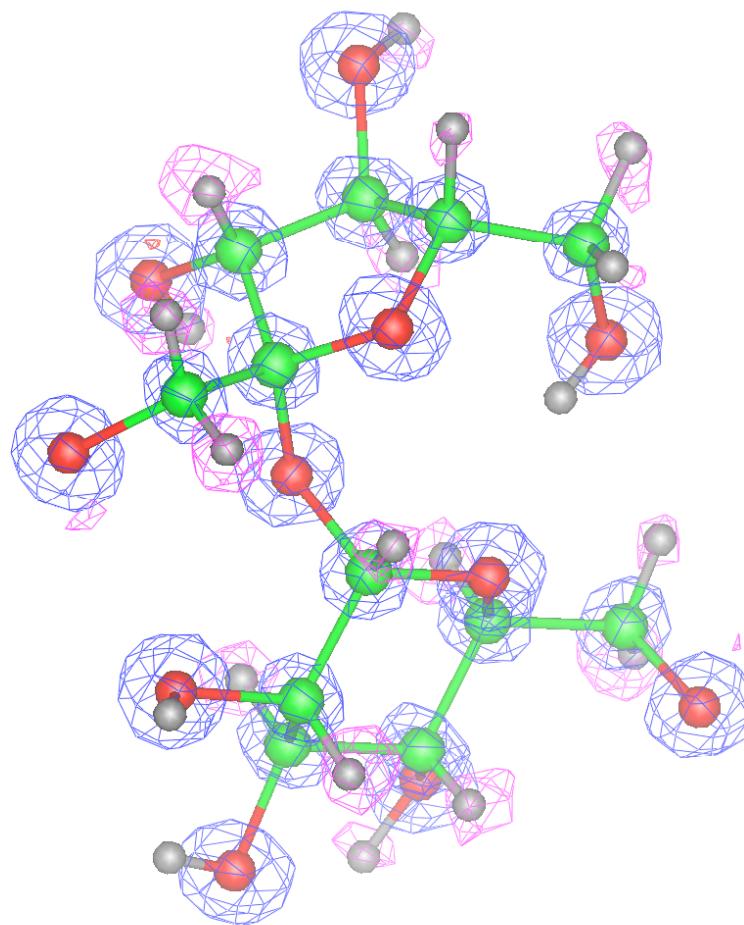
Blue: 2mFo-DFc

Pink: positive mFo-DFc

Sucrose (C&H)

ALS BL5.0.2

Refined with hydrogen contribution



Bias

- The phases dominate the looks of the image
- One should make sure that features in the density are not there because you put them there
 - Use Classic, SA or Full omit maps for confirmation
 - Omit map: remove a part of the structure and see if comes back in a difference map
 - SA: simulated annealing
 - Full omit map: includes density modification (PHENIX)



Software suites

- CCP4
 - <http://www.ccp4.ac.uk>
- CNS
 - <http://cns.csb.yale.edu/v1.2>
- **PHENIX**
 - <http://www.phenix-online.org>
- SHELX
 - <http://shelx.uni-ac.gwdg.de/SHELX>

Example Phenix applications

- Refinement
 - `phenix.refine mydata.sca mymodel.pdb`
- Structure solution
 - `phenix.autosol mydata.sca seq.txt`
- Twinned refinement
 - `phenix.refine mydata.sca mymodel.pdb twin_law="k,h,-l"`
- Data analyses
 - `Phenix.xtriage mydata.mtz`

Some pointers

- <http://www-structmed.cimr.cam.ac.uk/course.html>
 - Google on 'structural medicine course'
- Stout and Jensen; Drenth
- Molecular replacement basics
 - Crowther, R. A. and Blow, D. M. (1967) *Acta Crystallogr.* 23, 544-548.
 - Rossmann, M. G. and Blow, D. M. (1962). *Acta Cryst.* 15, 24-31.
- Density modification
 - Terwilliger, *Acta Cryst.*, (2000). D56, 965–972
- Refinement
 - G.N. Murshudov, A.A.Vagin and E.J.Dodson, (1997). *Acta Cryst.* D53, 240-255
- This talk
 - [**http://cci.lbl.gov/~phzward/Talks/SMB.pdf**](http://cci.lbl.gov/~phzward/Talks/SMB.pdf)

Acknowledgements

Gurussaakshaath param brahma tasmai shree gurave namaha

- **Henk Schenk**
- **Rene Peschar**
- Victor Lamzin
- Zbigniew Dauter
- Garib Murshudov
- Eleanor Dodson
- Tom Terwilliger
- Randy Read
- Gerard Bricgne
- Paul Adams
- Ralf Grosse-Kunstleve
- And many others